Mod B Live Lecture 1
6 May

Data Driven Healthcare
Xaim master

Difference of Proportions, Relative Risk, Odds Ratio

Measures useful for categorical data

- Difference of Proportions
- Relative Risk (Ratio of Proportions)
- Odds & Odds Ratio

Two Sample Problems for Proportions

Choose a SRS of size n_1 from a large population having proportion π_1 of successes and an independent SRS of size n_2 from another population having proportion π_2 of successes.

Population	Population proportion	Sample size	Count of successes	Estimate of π_i
1	π_1	n_1	y_1	$\widehat{\pi}_1 = y_1/n_1$
2	π_2	n_2	y_2	$\widehat{\pi}_2 = y_2/n_2$

Example: Physician's Health Study on Aspirin & Heart Attack

Physician's Health Study was a 5-year randomized study testing whether regular intake of aspirin reduces mortality from cardiovascular disease[1].

- Participants were males 40-84 years old in 1982 with no prior history of heart attack, stroke, and cancer, no current liver or renal disease, no contraindication of aspirin, no current use of aspirin
- Every other day, the males participating in the study took either one aspirin tablet or a placebo.
- Response: whether the participant had a heart attack (including fatal or non-fatal) during the 5 year period.

[1] *Source*: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physician's Health Study. *New Engl. J. Med.*, **318**: 262-64,1988.

Data: Physician's Health Study (Aspirin & Heart Attack)

Myocardial Infarction (MI) = heart attack. 2×2 table.

	MI	
Group	Yes	No
Placebo	189	10845
Aspirin	104	10933

Still 2×2:

	MI		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071

Difference of Proportions

Wald CI for Diff. of Proportions

Wald CI for $\pi_1 - \pi_2$ is

$$\widehat{\pi}_1 - \widehat{\pi}_2 \pm z_{\alpha/2}\sqrt{\frac{\widehat{\pi}_1(1-\widehat{\pi}_1)}{n_1} + \frac{\widehat{\pi}_2(1-\widehat{\pi}_2)}{n_2}}$$

Example: Aspirin & Heart Attack ---

	MI		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

$\Rightarrow \widehat{\pi}_1 = 189/11034 \approx 0.0171$

$\Rightarrow \widehat{\pi}_2 = 104/11037 \approx 0.0094$

95% CI for $\pi_1 - \pi_2$:

$$0.0171 - 0.0094 \pm 1.96\sqrt{\frac{0.0171 \times 0.9829}{11034} + \frac{0.009 \times 0.9906}{11037}}$$

$$= 0.0077 \pm 1.96(0.00154) = 0.0077 \pm 0.0030 = (0.0047, 0.0107)$$

Example: Aspirin & Heart Attack — Conclusion

- As the 95% CI does not contain 0, the incidence rate of heart attack was significantly lower in aspirin group than in the placebo group
- Can we claim that taking aspirin every other day is effective in reducing the chance of heart attack?

Example: Aspirin & Heart Attack — Conclusion

- As the 95% CI does not contain 0, the incidence rate of heart attack was significantly lower in aspirin group than in the placebo group
- Can we claim that taking aspirin every other day is effective in reducing the chance of heart attack?

Yes, because it was a randomized, double-blind, placebo-controlled experiment.

Agresti-Caffo Confidence Interval for $\pi_1 - \pi_2$

For small samples, Wald CI for $\pi_1 - \pi_2$ suffers from similar problem with achieving the nominal of confidence level as Wald CI for a single proportion.

Agresti and Caffo (2000) suggested adding *one success* and *one failure* in each of the two samples.

$$\tilde{\pi}_1 = \frac{y_1+1}{n_1+2}, \qquad \tilde{\pi}_2 = \frac{y_2+1}{n_2+2}$$

Agresti-Caffo CI for $\pi_1 - \pi_2$ is given by

$$(\tilde{\pi}_1 - \tilde{\pi}_2) \pm z_{\alpha/2} \sqrt{\frac{\tilde{\pi}_1(1-\tilde{\pi}_1)}{n_1+2} + \frac{\tilde{\pi}_2(1-\tilde{\pi}_2)}{n_2+2}}$$

Note we still estimate π_1 and π_2 by $\widehat{\pi}_1 = y_1/n_1$ and $\widehat{\pi}_2 = y_2/n_2$, not by $\tilde{\pi}_1$ and $\tilde{\pi}_2$.

The true confidence level of the Agresti-Caffo CI is closer to the nominal level than the Wald CI and hence is recommended.

Testing the Equality of Two Proportions

(Next Live lecture more on this)

The z-statistic for testing H_0: $\pi_1 = \pi_2$ is

$$z = \frac{\widehat{\pi}_1 - \widehat{\pi}_2}{\sqrt{\widehat{\pi}(1-\widehat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \text{where } \widehat{\pi} = \frac{y_1 + y_2}{n_1 + n_2}$$

Under H_0, z is approx. $N(0, 1)$.

Example: Aspirin & Heart Attack ----

Group	MI: Yes	MI: No	Total	
Placebo	189	10845	11034	$\Rightarrow \widehat{\pi}_1 = 189/11034 \approx 0.0171$
Aspirin	104	10933	11037	$\Rightarrow \widehat{\pi}_2 = 104/11037 \approx 0.0094$

For testing H_0: $\pi_1 = \pi_2$, $\quad \widehat{\pi} = \frac{189+104}{11034+11037} \approx 0.0132$

$$z = \frac{0.0171 - 0.0094}{\sqrt{0.0132(1-0.0132)\left(\frac{1}{11034} + \frac{1}{11037}\right)}} \approx \frac{0.0077}{0.00154} \approx 5.001$$

2-sided P-value = 0.00000057, strong evidence against H_0.

Small Sample Tests for 2×2 Tables

Note the test on the previous slide works for large sample only. Use it only when the numbers of successes and failures are both at least 5 in both samples (i.e., all n_{ij}'s are ≥ 5.)

		Success	Failure
Sample	1	n_{11}	n_{12}
	2	n_{21}	n_{22}

A small sample test for H_0: $\pi_1 = \pi_2$ is the Fisher's exact test)

Relative Risk

Why Ratio of Porportions?

E.g., consider the probability of a certain disease for smokers (π_1) and for nonsmokers (π_2):

- Case 1: $\pi_1 = 0.51$ and $\pi_2 = 0.50$
- Case 2: $\pi_1 = 0.011$ and $\pi_2 = 0.001$.

In both cases $\pi_1 - \pi_2 = 0.01$.

However, in Case 1, an increase of 0.01 due to smoking is small relative to the already sizable risk of disease in the nonsmoking population.

Case 2 has smokers with 11 times the chance of disease than nonsmokers.

When π_1 and π_2 are both small, the ratio π_1/π_2 might be a more relevant measure of the smoking effect than the difference $\pi_1 - \pi_2$.

Relative Risk (RR) = Ratio of Proportions

Relative Risk (RR) = $\frac{\pi_1}{\pi_2}$, estimated by = $\frac{\widehat{\pi}_1}{\widehat{\pi}_2}$.

Example: Aspirin & Heart Attack

Sample relative risk in the Physicians Health Study is

$$\frac{\widehat{\pi}_{placebo}}{\widehat{\pi}_{aspirin}} = \frac{0.0171}{0.0094} = 1.82$$

The risk (probability) of heart attacks was estimated to be 82% higher (or 1.82 times as high as) in placebo group than in the Aspirin group.

- Independence $\Longleftrightarrow \pi_1 = \pi_2 \Longleftrightarrow RR = \frac{\pi_1}{\pi_2} = 1$

Inference of Relative Risk (RR) π_1/π_2 (1)

- Sampling distribution for sample RR $(\widehat{\pi}_1/\widehat{\pi}_2)$ is highly skewed. The large sample normal approximation is NOT good.
- Sampling distribution of $\log(\widehat{\pi}_1/\widehat{\pi}_2)$ is closer to normal.
- It can be shown that

$$\mathrm{Var}\left(\log\left(\frac{\widehat{\pi}_1}{\widehat{\pi}_2}\right)\right) \approx \frac{1-\pi_1}{n_1\pi_1} + \frac{1-\pi_2}{n_2\pi_2}.$$

The SE of $\log(\widehat{\pi}_1/\widehat{\pi}_2)$ is thus

$$\mathrm{SE} = \sqrt{\widehat{\mathrm{Var}}\left(\log\left(\frac{\widehat{\pi}_1}{\widehat{\pi}_2}\right)\right)} = \sqrt{\frac{1-\widehat{\pi}_1}{n_1\widehat{\pi}_1} + \frac{1-\widehat{\pi}_2}{n_2\widehat{\pi}_2}}$$

$$= \sqrt{\frac{1}{y_1} - \frac{1}{n_1} + \frac{1}{y_2} - \frac{1}{n_2}}$$

Confidence Interval for Relative Risk (RR)

CI for $\log(\text{RR})$:

$$\log\left(\frac{\widehat{\pi}_1}{\widehat{\pi}_2}\right) \pm z_{\alpha/2}\text{SE} = \log\left(\frac{\widehat{\pi}_1}{\widehat{\pi}_2}\right) \pm z_{\alpha/2}\sqrt{\frac{1}{y_1} - \frac{1}{n_1} + \frac{1}{y_2} - \frac{1}{n_2}}$$
$$= (L, U)$$

CI for RR:

$$(e^L, e^U)$$

Example: Aspirin & Heart Attack

	MI		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

$\Rightarrow \widehat{\pi}_1 = 189/11034 \approx 0.0171$

$\Rightarrow \widehat{\pi}_2 = 104/11037 \approx 0.0094$

SE for $\log(\text{RR})$ is

$$\sqrt{\frac{1}{y_1} - \frac{1}{n_1} + \frac{1}{y_2} - \frac{1}{n_2}} = \sqrt{\frac{1}{189} - \frac{1}{11034} + \frac{1}{104} - \frac{1}{11037}} \approx 0.1213$$

95% CI for $\log(\text{RR})$ is

$$\begin{aligned}\log(\widehat{\pi}_1/\widehat{\pi}_2) \pm z_{\alpha/2}\text{SE} = \log\left(\frac{0.0171}{0.0094}\right) \pm 1.96(0.1213) &= 0.5984 \pm 0.2378 \\ &\approx (0.3606, 0.8362).\end{aligned}$$

95% CI for RR is $(e^{0.3606}, e^{0.8362}) = (1.4342, 2.3076)$.

Interpretation. With 95% confidence, the risk of MI with 5 years for male physicians taking placebo is 1.43 to 2.30 times as high (or 43% to 130% higher) compared to those taking Aspirin.

Interpretation of Relative Risk

		Response (Y)	
		Success	Failure
Explanatory	Group 1	π_1	$1 - \pi_1$
Variable (X)	Group 2	π_2	$1 - \pi_2$

Interpretation of the Relative Risk (RR) $= \pi_1/\pi_2$:

The probability (or risk) of success is RR times as large in group 1 as in group 2.

Interpretation of a 95% CI (L, U) for the RR :

With 95% confidence, the probability (or risk) of success is L to U times as large in group 1 as in group 2.

Here, “success”, “group 1”, and “group 2” would be replaced with meaningful terms in context.

Example

A 20-year study of British male physicians noted that the proportion who died from heart disease was

- $\widehat{\pi}_1 = 0.00669$ for smokers and
- $\widehat{\pi}_2 = 0.00413$ for nonsmokers.

Interpretation of (estimated) Relative Risk (RR) $= \widehat{\pi}_1/\widehat{\pi}_2 = 1.62$:

> *The estimated probability (or risk) of dying from heart disease for cigarette smokers was 1.62 times as large compared to nonsmokers.*

or

> *Cigarette smokers were estimated to be 1.62 times as likely as (or 62% more likely than) nonsmokers to die from heart disease*

Interpretation of Relative Risk

- If RR < 1, it might be more appealing to say
  The probability/risk of success is $1/RR$ times as large in group 2 as in group 1.

than to say

The probability/risk of success is RR times as large in group 1 as in group 2.

since "11 times as large" might sound more striking than "$1/11 = 0.091$ times as large."

- However, it might be okay to keep RR < 1 if Group 1 is some treatment and Group 2 is a placebo since in that case RR < 1 means reduction of risk.
- To sum up, we may, but not always have to, make RR > 1

Side Notes: Vaccine Efficacy & Relative Risk

Vaccine efficacy/effectiveness is defined to be

$$1 - RR = 1 - \frac{\pi_{vaccinated}}{\pi_{unvaccinated}}.$$

Side Notes: Vaccine Efficacy & Relative Risk

Vaccine efficacy/effectiveness is defined to be

$$1 - RR = 1 - \frac{\pi_{vaccinated}}{\pi_{unvaccinated}}.$$

e.g., 2 dose of Pfizer vaccine + booster had 70% efficacy immediately after the third dose

- 70% efficacy doesn't mean you have a 30% chance of getting sick
- 70% efficacy $\Rightarrow$ RR = 0.3, the chance of getting sick for vaccinated people is only 30% of the chance for unvaccinated people

Odds & Odds Ratio

Odds

Consider a variable with binary outcome {Success, Failure}={S, F} (or {Yes, No})

	Outcome	
	Success	Failure
probability	π	$1-\pi$

The odds of outcome S (instead of F) is

$$odds(S) = \frac{\mathrm{P}(S)}{\mathrm{P}(F)} = \frac{\pi}{1-\pi}.$$

- if odds $= 3$, then S is three times as likely as F;
- if odds $= 1/3$, then F is three times as likely as S.

$$\mathrm{P}(S) = \pi = \frac{odds(S)}{1 + odds(S)}$$

$$odds(S) = 3 \Longrightarrow \mathrm{P}(S) = \frac{3}{1+3} = \frac{3}{4}, \qquad \mathrm{P}(F) = \frac{1}{4}$$

$$odds(S) = \frac{1}{3} \Longrightarrow \mathrm{P}(S) = \frac{1/3}{1+1/3} = \frac{1}{4}, \quad \mathrm{P}(F) = \frac{3}{4}$$

Odds Ratio (OR) = Ratio of Odds

Population	Population proportion	Sample size	Count of successes	Estimate of π_i
1	π_1	n_1	y_1	$\widehat{\pi}_1 = y_1/n_1$
2	π_2	n_2	y_2	$\widehat{\pi}_2 = y_2/n_2$

$$odds(\text{Success}) = \begin{cases} \dfrac{\pi_1}{1-\pi_1} & \text{in population 1} \\ \dfrac{\pi_2}{1-\pi_2} & \text{in population 2} \end{cases}$$

Definition (Odds Ratio)

$$Odds\ Ratio: \quad \theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

Relative Risk (RR) v.s. Odds Ratio (OR

$$\text{Odds Ratio} = \text{Relative Risk} \times \frac{1 - \pi_2}{1 - \pi_1}$$

When $\pi_1 \approx 0$ and $\pi_2 \approx 0$,

$$\text{Odds Ratio} \approx \text{Relative Risk}$$

Odds ratio is more further away from 1 than relative risk (RR)

- If $\pi_1 > \pi_2$, then Odds Ratio $>$ RR > 1.
- If $\pi_1 < \pi_2$, then Odds Ratio $<$ RR < 1.

Estimate of Odds Ratio

		Success	Failure	Total
Population	1	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+2}	n_{++}

$$\widehat{\theta} = \frac{\widehat{\pi}_1/(1-\widehat{\pi}_1)}{\widehat{\pi}_2/(1-\widehat{\pi}_2)} = \frac{(n_{11}/n_{1+})/(n_{12}/n_{1+})}{(n_{21}/n_{2+})/(n_{22}/n_{2+})} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Odds ratio is thus called the **cross-product ratio**.

Properties of the Odds Ratio

- odds > 0, $\theta > 0$
- $\theta = 1$ when $\pi_1 = \pi_2$; i.e., when X, Y are independent.
- The further θ is away from 1, the stronger the association.

  e.g.,For Y = lung cancer, studies estimated that

  - $\theta \approx 10$ for X = smoking,
  - $\theta \approx 2$ for X = passive smoking.

- If rows are interchanged (or if columns are interchanged),

$$\theta \longrightarrow 1/\theta.$$

  e.g., a value of $\theta = 1/5$ indicates the same strength of association as $\theta = 5$, but in the opposite direction.

Log Odds Ratio

- Sampling distribution of $\widehat{\theta}$ is skewed to the right.
  Normal approximation for $\widehat{\theta}$ is NOT good.

Log Odds Ratio

- Sampling distribution of $\widehat{\theta}$ is skewed to the right.
  Normal approximation for $\widehat{\theta}$ is NOT good.
- Sampling distribution of $\log \widehat{\theta}$ is closer to normal.

Log Odds Ratio

- Sampling distribution of $\widehat{\theta}$ is skewed to the right.
  Normal approximation for $\widehat{\theta}$ is NOT good.
- Sampling distribution of $\log \widehat{\theta}$ is closer to normal.
- $\theta = 1 \iff \log \theta = 0$, when X, Y are independent
- If rows (or columns) are interchanged,

$$\log \theta \longrightarrow \log(1/\theta) = -\log \theta.$$

The log odds ratio ($\log \theta$) is symmetric about 0, e.g.,

$$\theta = 2 \iff \log \theta = 0.7$$
$$\theta = 1/2 \iff \log \theta = -0.7$$

The absolute value of $\log \theta$ indicates the strength of association

A Confidence Interval for the Odds Ratio

Large-sample (asymptotic) SE of $\log \widehat{\theta}$ is

$$\mathrm{SE}(\log \widehat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

A Confidence Interval for the Odds Ratio

Large-sample (asymptotic) SE of $\log\widehat{\theta}$ is

$$\mathrm{SE}(\log\widehat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

CI for $\log\theta$:

$$(L, U) = \log\widehat{\theta} \pm z_{\alpha/2} \times \mathrm{SE}(\log\widehat{\theta})$$

A Confidence Interval for the Odds Ratio

Large-sample (asymptotic) SE of $\log \widehat{\theta}$ is

$$\mathrm{SE}(\log \widehat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

CI for $\log \theta$:

$$(L, U) = \log \widehat{\theta} \pm z_{\alpha/2} \times \mathrm{SE}(\log \widehat{\theta})$$

CI for θ:

$$(e^L, e^U).$$

Example (Aspirin & Heart Attack)

	MI	
Group	Yes	No
Placebo	189	10845
Aspirin	104	10933

$$\widehat{\theta} = \frac{189 \times 10933}{104 \times 10845} = 1.83$$

$$\log \widehat{\theta} = \log(1.83) = 0.605$$

$$\mathrm{SE}(\log \widehat{\theta}) = \sqrt{\frac{1}{189} + \frac{1}{10845} + \frac{1}{104} + \frac{1}{10933}} = 0.123$$

95% CI for $\log \theta : 0.605 \pm 1.96(0.123) = (0.365, 0.846)$

95% CI for $\theta : (e^{0.365}, e^{0.846}) = (1.44, 2.33)$

Remarks

- Apparently $\theta > 1$.
- $\widehat{\theta}$ is not at the midpoint of the CI
- Better estimate if we use $\{n_{ij} + 0.5\}$, specially if any $n_{ij} = 0$.